Journal of Agricultural Sciences (JAS) 32(2):423-438
DOI: 10.15832/ankutbd.1809955

J Agr Sci-Tarim Bili
e-ISSN: 2148-9297
jas.ankara.edu.tr

Multi-Depth Soil Moisture Prediction Using Machine Learning Across Türkiye's Diverse Environments

Muhammed Sungur Demir[a*]

[a]*Civil Engineering Department, Faculty of Engineering and Architecture, Batman University, 72100, TÜRKİYE*

ARTICLE INFO
Research Article
Corresponding Author: Muhammed Sungur Demir, E-mail: muhammed.sungur@batman.edu.tr
Received: 24 October 2025 / Revised: 18 December 2025 / Accepted: 29 December 2025 / Online: 24 March 2026

Cite this article
Demir M S (2026). Multi-Depth Soil Moisture Prediction Using Machine Learning Across Türkiye's Diverse Environments. *Journal of Agricultural Sciences (Tarim Bilimleri Dergisi),* 32(2):423-438. DOI: 10.15832/ankutbd.1809955

ABSTRACT
Efficient agricultural water management is essential to address global water scarcity. Soil moisture serves as a key indicator for irrigation scheduling, crop yield forecasting, and hydrological modeling. However, traditional in-situ measurements are costly, labor-intensive, and spatially limited. This study develops a machine learning framework for predicting soil moisture at multiple depths (20, 40, and 80 cm) using readily available environmental variables. Türkiye was selected as the study area due to its pronounced climatic and pedological heterogeneity, providing an ideal testbed for training robust models. Seven machine learning algorithms were systematically evaluated using data from 201 meteorological stations spanning 2016–2024. Extreme Gradient Boosting (XGBoost) outperformed all alternatives following hyperparameter optimization, achieving strong accuracy across all depths ($R^2 = 0.74$, 0.69, and 0.66 at 20, 40, and 80 cm, respectively). Feature importance analysis indicated a depth-dependent shift in the controls on soil moisture. Near the surface (20 cm), short-term meteorological variability and seasonal dynamics were the primary drivers, whereas deeper layers were increasingly regulated by stable soil hydraulic properties, with clay dominating at intermediate depths, and large-scale spatial gradients represented by latitude and elevation. This pattern reflects a transition from weather-driven processes at shallow depths to long-term regulation by soil structure and regional climate patterns in the subsurface. The model maintained robust performance across diverse environmental conditions, with seasonal accuracy varying by nearly 10% between winter and autumn. Spatial analysis revealed regional variations in controlling factors: soil properties (particularly organic carbon) dominated in the northern regions, while their relative importance decreased southward where climatic and temporal variables contributed more substantially, reflecting heightened sensitivity of moisture dynamics to meteorological fluctuations. The framework provides a scalable, cost-effective solution for soil moisture monitoring in data-scarce regions, supporting irrigation optimization, drought early warning, and sustainable water governance.

Keywords: XGBoost, Root-zone soil moisture, Machine learning model comparison, Temporal cross-validation, Agricultural water management

1. Introduction

Soil moisture is a critical component of the hydrological cycle, yet it remains one of the least directly measured hydrological variables despite its fundamental importance (Harani et al. 2025; Sun et al. 2025). It plays an essential role in plant water and nutrient uptake, and its dynamics directly influence surface runoff generation and deep percolation processes (Mdemu et al. 2020; Harani et al. 2025). Specifically, following precipitation events, water is partitioned into several pathways: a portion becomes surface runoff, some evaporates back to the atmosphere without infiltrating the soil, and the remainder infiltrates into the soil, increasing soil moisture content. When sufficient moisture accumulates, percolation to deeper layers occurs (Bayazıt 2013). Soil moisture also serves as a key indicator of climate change impacts on terrestrial water cycles and plays a decisive role in flood generation and groundwater recharge dynamics (Chen et al. 2025).

In the agricultural sector, soil moisture is equally vital. It is particularly crucial for predicting crop yields in rainfed agricultural systems (Chen et al. 2025), where temporal variability in soil moisture directly influences productivity (Champagne et al. 2019). Both water deficit and excess adversely affect plant performance: insufficient moisture limits plant growth and causes yield losses, while over-irrigation restricts root oxygen availability and nutrient uptake, also reducing productivity (Seleiman et al. 2021; Irmak 2014). Maintaining this delicate balance requires accurate soil moisture monitoring to enable optimal irrigation scheduling and frequency, thereby improving water-use efficiency and mitigating these adverse effects (Mdemu et al. 2020). Furthermore, soil moisture data support drought and flood early-warning systems (Teshome et al. 2024), underscoring its central role in sustainable agricultural water management (Chen et al. 2025).

Despite its significance, accurately predicting soil moisture remains challenging due to the complex and nonlinear interactions among hydrological, atmospheric, and soil processes (Wang et al. 2024). Soil moisture dynamics are governed by multiple factors including precipitation patterns, evapotranspiration rates, soil hydraulic properties, vegetation characteristics, and

topographic features, all of which exhibit high spatial and temporal heterogeneity (Rasheed et al. 2022). This complexity makes soil moisture one of the most difficult hydrological variables to estimate across large areas (Karthikeyan & Mishra 2021).

Traditional approaches to soil moisture estimation rely on in-situ measurements, which are generally classified into direct and indirect methods (Rasheed et al. 2022). The gravimetric method, which involves oven-drying soil samples to determine water content, is considered the benchmark for accuracy (Susha Lekshmi et al. 2014). However, it is labor-intensive, time-consuming, and destructive to sampling sites, limiting its application for continuous monitoring (Walker et al. 2004; Zha et al. 2025). Indirect methods estimate soil moisture by measuring physical properties affected by water content, such as dielectric permittivity or electrical conductivity. While these techniques avoid destructive sampling, they are often costly, require soil-specific calibration, and can produce errors in saline environments (Susha Lekshmi et al. 2014). Radiometric techniques such as neutron scattering face constraints related to health risks and operational complexity (Harani et al. 2025), whereas cosmic-ray neutron sensors, despite recent advances, remain prohibitively expensive for widespread deployment (Champagne et al. 2019). Moreover, all ground-based measurements are inherently point-based, which limits their representativeness in spatially heterogeneous landscapes (Karthikeyan & Mishra 2021).

To overcome these spatial limitations, remote sensing techniques have gained prominence over the past decades. These approaches utilize optical, thermal, or microwave satellite observations across multiple spectral bands to estimate soil moisture over large areas (Taktikou et al. 2016). The primary advantage of remote sensing lies in its capacity to provide spatially continuous and temporally repetitive measurements in a cost-effective manner, making it particularly suitable for agricultural and hydrological applications (Nichols 2011). However, several limitations constrain its operational use. Vegetation cover, surface roughness, and topographic effects can significantly reduce retrieval accuracy (Nichols 2011). Moreover, most satellite products offer limited spatial and temporal resolution (Champagne et al. 2019). Furthermore, remote sensing techniques are generally sensitive only to near-surface soil layers, typically the top 5 cm (Zha et al. 2025), which may not adequately represent root-zone moisture conditions critical for agricultural applications. To address some of these limitations, integrating active and passive microwave observations can enhance accuracy; however, such integration introduces greater analytical and computational complexity (Ahmad et al. 2010).

More recently, advances in artificial intelligence and machine learning have opened new avenues for soil moisture estimation, demonstrating high predictive accuracy with reduced costs and computational time. Various algorithms have been systematically evaluated across diverse environmental conditions. Houben et al. (2025) compared Random Forest (RF), Gradient Boosted Regression Trees (GBRT), Support Vector Regression (SVR), and Neural Networks (NN) using dense sensor networks in Germany, while Teshome et al. (2024) evaluated machine learning and deep learning approaches in the United States. Zha et al. (2025) integrated Random Forest models with remote sensing data in an arid region of northern China. For multi-depth prediction, Chen et al. (2025) and Zhang et al. (2025) assessed model performance across multiple soil layers. At larger spatial scales, Wang et al. (2024) explored deep learning models for global soil moisture prediction. Regional studies (Nguyen et al. 2022; Zhu et al. 2025) have further demonstrated the versatility of machine learning approaches under varying climatic and soil conditions.

However, despite this extensive international research, no comprehensive machine learning-based study has been conducted in Türkiye. The country presents unique opportunities for such research due to its vast geographic extent, heterogeneous soil structure, and pronounced climatic and topographic variability. These characteristics enable model training across a wide range of environmental conditions, potentially enhancing predictive robustness and generalization. Moreover, given Türkiye's large population, intensive agricultural production, and growing water stress, developing cost-effective and accurate soil moisture prediction tools holds substantial practical importance for agricultural water management.

The main objective of this study is to develop and evaluate machine learning models for predicting soil moisture at multiple depths (20, 40, and 80 cm) across Türkiye using climatic, topographic, and soil-related variables. Specifically, the study aims to: (1) compare the performance of different machine learning algorithms to identify suitable approaches for soil moisture prediction, and (2) assess how the relative importance of predictors varies across different depths, seasons, and regions. The proposed framework is intended to generate soil moisture estimates in areas lacking in-situ observations, providing a practical tool for irrigation scheduling, crop yield forecasting, and drought assessment. The findings may contribute to improved water resource management and support agricultural planning in Türkiye.

2. Material and Methods

2.1. Study area

Türkiye is located in the eastern Mediterranean region, covering an area of approximately 783,000 km² between 36°–42° N latitude and 26°–45° E longitude. The country exhibits diverse climatic conditions due to its large geographical extent, strategic location at the intersection of different climate zones, and rugged topography with elevations ranging from sea level to over 5,000 m (Yetik et al. 2024). This geographic diversity creates distinct regional climate patterns. The country is predominantly influenced by Mediterranean climate along its coastal regions, while its southeastern areas adjacent to desert climate zones

experience arid conditions (Özütürk 2024), and the regions along the Black Sea coast are characterized by high precipitation throughout the year (Cengiz et al. 2022).

The spatial distributions of mean temperature, precipitation, and soil moisture during the study period are presented in Figure 1. The country receives an average annual precipitation of approximately 573 mm (MGM 2023), but this distribution is highly heterogeneous across regions. Annual precipitation reaches 2,500 mm in the eastern Black Sea region, while it drops to as low as 250 mm in some areas (Ölgen 2010), reflecting the strong influence of topography and proximity to moisture sources.

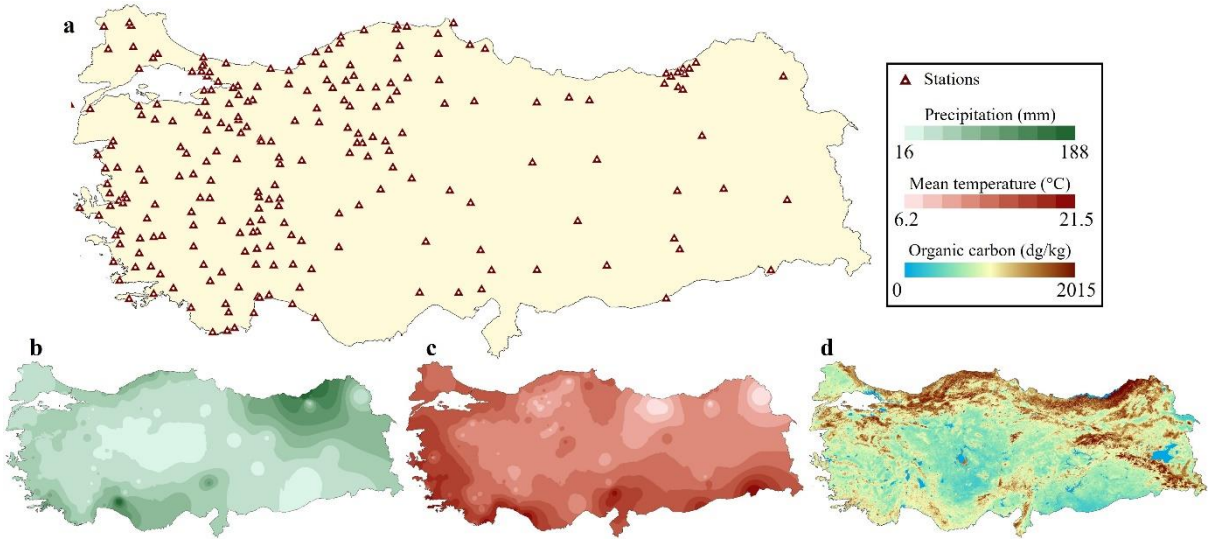


Figure 1- Spatial distribution of meteorological stations and environmental variables across Türkiye: (a) station locations (n=201), (b) mean annual precipitation (mm), (c) mean annual temperature (°C), and (d) soil organic carbon content (dg/kg) in the 0–5 cm layer

The mean annual temperature in Türkiye is 13.3 °C, with substantial seasonal and spatial variability. Seasonally, the country-wide average temperature decreases to 2.4 °C in January and rises to 24.5 °C in July. Spatially, mean annual temperatures reach 20.2 °C in the southern regions, while they decline significantly to approximately 3.8 °C toward the northeastern highlands (MGM 2024).

The spatial distribution of soil moisture, based on meteorological station observations, is also shown in Figure 1. Soil moisture is relatively higher in the northern and western parts of the country, corresponding to regions with greater precipitation inputs and lower evaporative demand. Conversely, soil moisture is notably lower in the southeastern region, where limited precipitation and high evapotranspiration rates constrain water availability.

Figure 1 also presents the distribution of soil organic carbon content in the 0–5 cm depth layer, which was identified as an important parameter for soil moisture prediction in the machine learning models. Organic carbon content is particularly high along the Black Sea coast and in eastern Anatolia, where humid climatic conditions and dense vegetation cover promote organic matter accumulation. In contrast, organic carbon levels are considerably lower in the central Anatolia region, especially around the Salt Lake area, where arid conditions and sparse vegetation limit soil organic matter development.

Türkiye’s sharp climatic gradients, diverse soil types, and complex topography make it an ideal natural laboratory for testing soil moisture prediction across contrasting environments.

2.2. Data and preprocessing

This study utilized a soil moisture dataset collected from 201 meteorological stations across Türkiye between 2016 and 2024 (Figure 1). The dataset includes soil moisture measurements at three depths: 20, 40, and 80 cm. However, data availability was inconsistent across stations, depths, and time periods. Some stations lacked observations for certain years or months, while measurements at deeper layers (40 and 80 cm) were particularly limited. Consequently, the dataset comprised 17,905 observations for the 20 cm depth, and 2,280 observations for each of the 40 cm and 80 cm depths. In addition to these temporal and depth-related gaps, the spatial distribution of stations was also uneven, with denser coverage in western and coastal regions and sparse representation in central and southeastern Türkiye.

To predict soil moisture variations across diverse conditions, eighteen variables representing climatic, topographic, and soil characteristics were compiled as potential predictors. These included Year, Month, Latitude (Lat), Longitude (Lon), Elevation (E), Precipitation (P), Relative Humidity (RH), Wind Speed (WS), Mean/Maximum/Minimum Temperature (T_{mean}, T_{max} and

T_{min}), Solar Radiation (SR), and soil properties measured at 0–5 cm depth: Bulk Density (BD), Sand/Silt/Clay contents, Organic Carbon (OC), and Coarse Fragment (CF). For stations lacking solar radiation measurements, radiation values were estimated from recorded sunshine duration using the empirical approach described in FAO-56 (Allen et al. 1998) (Equation 1).

$$R_s = \left(a_s + b_s \frac{n}{N} \right) R_a \quad (1)$$

Where; R_s is solar radiation (MJ m^{-2} day^{-1}), n is actual sunshine duration (hours), N is maximum possible sunshine duration (hours), R_a is extraterrestrial radiation (MJ m^{-2} day^{-1}), and a_s and b_s are empirical coefficients (recommended values: a_s= 0.25, b_s = 0.50).

To reduce potential redundancy among this relatively large set of variables, a comprehensive multi-criteria feature selection procedure was applied. This approach integrated Pearson correlation analysis to identify linear relationships among predictors, Variance Inflation Factor (VIF) assessment to quantify multicollinearity, and three feature importance methods: Random Forest (impurity-based), Permutation Importance, and SHAP (SHapley Additive exPlanations) values. Variable pairs with high correlation ($|r| > 0.80$) were identified as potentially redundant. Variables with consistently low importance across all methods and those exhibiting severe multicollinearity (supported by both high correlation and elevated VIF values) were excluded. Certain predictors with high VIF were retained when they represented distinct physical processes, as confirmed by consistent importance across multiple methods (detailed justification in Section 3.2). As a result, six variables (T_{max}, T_{min}, Sand and Silt contents, CF and SR) were eliminated, yielding a final set of 12 predictors for all subsequent analyses.

To eliminate scale-dependent effects and ensure compatibility across machine learning algorithms, all predictor variables were standardized using Z-score normalization (Equation 2).

$$z_i = \frac{x_i - \mu}{\sigma} \quad (2)$$

Where; x_i is the original value, μ is the mean, and σ is the standard deviation. After standardization, to prevent temporal information leakage and ensure realistic model evaluation, observations were partitioned chronologically rather than randomly: training set (2016–2021: 65%, n = 11,694 for 20 cm depth), and test set (2022–2024: 35%, n = 6,211 for 20 cm depth). This temporal split ensures that the model is trained exclusively on historical data and tested on future observations, mirroring operational conditions and avoiding artificially inflated performance metrics that can arise from random splitting of time-series data.

To ensure reproducibility, a fixed random seed (seed = 42) was applied to all stochastic processes within the machine learning algorithms, including bootstrap sampling, tree construction, and neural network weight initialization.

2.3. Machine learning methodology

This study employed a structured machine learning framework to predict soil moisture using meteorological, topographic, and soil-related variables. The overall workflow is illustrated in Figure 2, summarizing the sequential steps from data collection and feature selection to model training, evaluation, and interpretation.

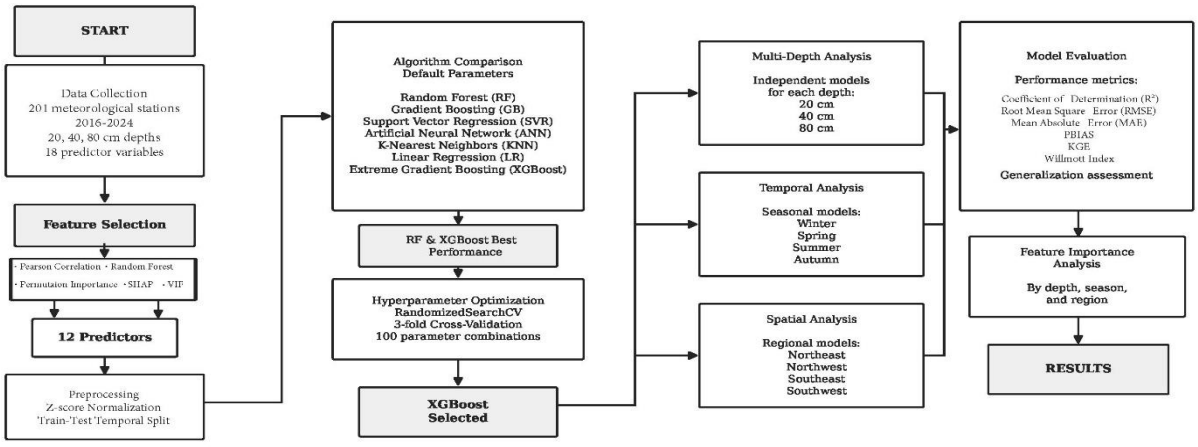


Figure 2- Machine learning workflow: data preprocessing, feature selection, temporal train-test split (2016–2021 vs 2022–2024), model training and hyperparameter optimization, and multi-metric performance evaluation

2.3.1. Algorithm selection and comparative framework

Seven machine learning algorithms were systematically evaluated for soil moisture prediction: RF, Gradient Boosting (GB), SVR, Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), Linear Regression (LR), and XGBoost. The modeling workflow was structured in two sequential stages to identify the most suitable approach. In the first stage, all algorithms were trained using default parameter settings (Table S1, Supplementary Material) to establish a baseline performance comparison. This preliminary assessment enabled the identification of models demonstrating superior predictive capability. Based on comparative results, RF and XGBoost emerged as the best-performing algorithms and were subsequently selected for detailed optimization and in-depth analysis.

2.3.2. Hyperparameter optimization and algorithm selection

Comprehensive hyperparameter tuning was performed for both RF and XGBoost at the 20 cm depth to identify the algorithm best suited for subsequent modeling across all soil depths. For XGBoost, optimization was conducted using RandomizedSearchCV with 5-fold TimeSeriesSplit cross-validation, evaluating 100 distinct parameter combinations. TimeSeriesSplit was specifically employed to preserve temporal order during cross-validation, ensuring that each fold trains on past data and validates on subsequent periods, thereby preventing temporal information leakage.

Following optimization, RF showed minimal performance change and was retained with default configuration, while XGBoost (optimized parameters in Table S2, Supplementary Material) maintained comparable test accuracy but demonstrated substantially reduced overfitting. Consequently, XGBoost was selected as the primary algorithm for all subsequent analyses across all soil depths.

2.3.3. Extreme gradient boosting (XGBoost)

XGBoost builds an ensemble of decision trees in sequence. The algorithm constructs trees sequentially, where each subsequent tree is optimized to minimize the residual errors from preceding trees, resulting in progressive improvement of predictive accuracy (Chen & Guestrin 2016; Romero Martínez et al. 2025). A major advantage of XGBoost lies in its integrated regularization mechanism within the objective function, which prevents overfitting and controls model complexity (Szczepanek 2022). Moreover, the algorithm employs a second-order Taylor approximation for loss function optimization, enabling more efficient and stable convergence than first-order gradient-based approaches (Ma et al. 2021; Wang et al. 2022). By combining regularization with iterative learning, XGBoost effectively manages multicollinearity among predictors and captures nonlinear interactions (Ali et al. 2023; Li et al. 2025), making it particularly suitable for hydrological and environmental applications (Guo et al. 2025).

2.3.4. Model evaluation and validation

Model performance was quantified using three complementary statistical metrics. Coefficient of Determination (R^2) measures the proportion of variance in observed soil moisture explained by model predictions (Equation 3).

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (x_i - x'_i)^2}{\sum_{i=1}^{n} (x_i - \overline{x}_i)^2} \quad (3)$$

Where; x_i is the observed value, x'_i is the predicted value, and $\bar{y}$ the mean of observed values. Values range from 0 to 1, with higher values indicating superior fit. Root Mean Square Error (RMSE) represents the standard deviation of prediction errors (Equation 4).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(x_i - x'_i)^2}{n}} \quad (4)$$

expressed as a percentage of mean observed soil moisture to facilitate scale-independent comparison. Mean Absolute Error (MAE) captures the average magnitude of prediction errors (Equation 5).

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |x_i - x'_i| \quad (5)$$

providing a direct, interpretable measure of model accuracy independent of error direction. Together, these complementary indicators provide comprehensive evaluation of model fit and predictive reliability.

2.3.5. Generalization assessment and data leakage prevention

To ensure scientific rigor and unbiased performance evaluation, a strict validation protocol was consistently applied across all soil depths (20, 40, and 80 cm). Hyperparameter optimization was performed exclusively on the training subset (2016-2021) using 5-fold TimeSeriesSplit cross-validation, while the independent test set (2022-2024) was reserved solely for final evaluation. This time-preserving validation strategy ensures that each fold uses only preceding observations for training and subsequent observations for validation, preventing temporal leakage and maintaining chronological integrity throughout the cross-validation process.

Model robustness and generalization were verified through two complementary criteria. The Training–Test R^2 Difference was calculated as:

$$\Delta R^2 = R^2_{train} - R^2_{test} \tag{6}$$

Where; values below 0.10 indicated strong generalization, 0.10–0.15 acceptable, and larger differences potential overfitting. Additionally, cross-validation–test consistency was assessed using:

$$\left| R^2_{cv} - R^2_{test} \right| < 0.1 \tag{7}$$

Where; the cross-validation R^2 is calculated as:

$$R^2_{cv} = \frac{1}{k} \sum_{i=1}^{k} R^2_i \tag{8}$$

This metric ensures agreement between cross-validation and test performance. Larger discrepancies in either criterion were interpreted as potential indicators of overfitting or information leakage. This comprehensive validation framework ensured that all reported model results reflect genuine predictive capability rather than artifacts of sampling bias or overfitting.

2.3.6. Temporal and spatial variability analysis

To investigate the temporal and spatial heterogeneity of soil moisture dynamics, sub-group analyses were conducted based on seasonal and geographic stratification. Four seasons were defined as winter (December to February), spring (March to May), summer (June to August), and autumn (September to November). For spatial segmentation, Türkiye was divided into four geographic regions: Northeast, Northwest, Southeast, and Southwest, based on the median latitude and longitude of monitoring stations.

Separate XGBoost models were trained for each depth and region. Feature importance values are reported as normalized percentages. At a depth of 20 cm, each seasonal group contained approximately 5,000 observations, whereas at 40 cm and 80 cm depths, each group comprised roughly 570 observations. The substantial reduction in data availability at deeper layers limited the robustness of stratified modeling; therefore, detailed assessments of temporal and spatial variability were restricted to the 20 cm depth.

3. Results and Discussion

3.1. Descriptive statistics of soil moisture

Descriptive statistics of soil moisture across the three depths are presented in Table 1. The surface layer (20 cm) contained 17,905 observations, whereas the 40 cm and 80 cm layers were limited to 2,280 observations each. This disparity reflects the available data from the meteorological monitoring network, where subsurface measurements were less extensively recorded during the study period.

Mean soil moisture was higher at the surface (20 cm: 21.15%) than at subsurface depths (40 cm: 14.45%; 80 cm: 15.29%). All three distributions exhibited positive skewness (0.89–1.02), indicating that soil moisture values typically cluster at lower levels with occasional increases following precipitation events.

Kurtosis values decreased with depth, from 1.20 at the surface to 0.37 at 40 cm and 0.36 at 80 cm. The higher kurtosis at the surface reflects a more peaked distribution with occasional extreme values, consistent with rapid wetting-drying cycles. Subsurface layers exhibited lower kurtosis, indicating more uniform moisture conditions.

Seasonal coefficient of variation (CV) ranged from 0.51 in winter to 0.71 in summer and autumn at 20 cm depth, reflecting greater moisture fluctuations during periods of high evapotranspiration demand and irregular precipitation. At deeper layers, CV values were substantially higher (40 cm: 0.96–1.27; 80 cm: 0.97–1.23). This elevated CV primarily reflects the lower mean moisture content at depth rather than greater absolute variability, though interpretation also requires caution given that subsurface sensors were not available at all stations. These distinct distributional characteristics motivated the adoption of independent, depth-specific modeling approaches.

Table 1- Descriptive statistics and seasonal variability of observed volumetric soil moisture (%) at 20 cm, 40 cm, and 80 cm depths

Statistic	*20 cm*	*40 cm*	*80 cm*
Sample Size (n)	17,905	2,280	2,280
Mean (%)	21.15	14.45	15.29
Std. Dev. (%)	13.4	15.73	16.51
Min – Max (%)	0.00–89.00	0.00–71.60	0.00–73.80
IQR (%)	17.4	26.9	25.5
Skewness	0.89	0.98	1.02
Kurtosis	1.2	0.37	0.36
Seasonal Variance (CV)*	0.51–0.71	0.96–1.27	0.97–1.23

*: Coefficient of variation (CV) range: winter (minimum) to summer (maximum)

3.2. Correlation analysis and feature selection

To ensure robust predictor selection and address the limitations of standard correlation methods, which may overlook nonlinear relationships or exhibit bias toward high-cardinality features, a multi-criteria approach was implemented. This framework integrated Pearson correlation, Variance Inflation Factor (VIF) assessment, and three distinct feature importance metrics: Random Forest (impurity-based), Permutation Importance, and SHAP values.

Pearson correlation analysis revealed that RH exhibited the strongest linear association with soil moisture ($r = 0.33$), followed by latitude ($r = 0.28$) and precipitation ($r = 0.25$) (Figure 3). Conversely, several variables showed negligible linear relationships, including Year ($r = -0.06$), Longitude ($r = 0.02$), and Wind Speed ($r = -0.08$). However, relying solely on linear correlation proved insufficient for capturing complex hydrological interactions.

A comparative analysis of feature importance techniques revealed substantial divergences in predictor rankings, confirming the necessity of a multi-view approach (Figure 3). While linear methods undervalued spatiotemporal variables, the Random Forest, Permutation, and SHAP analyses consistently identified Latitude as a dominant predictor (RF: 14.6%, Permutation: 20.1%, SHAP: 16.2%), despite its moderate linear correlation. Longitude also demonstrated notable importance across tree-based methods (RF: 9.7%, Permutation: 10.2%, SHAP: 6.7%), reflecting the west–east continentality gradient across Türkiye. Year exhibited modest importance (RF: 3.0%, Permutation: 3.2%, SHAP: 2.35%), contrasting with its negligible linear correlation ($r = -0.06$). This suggests the presence of subtle temporal trends in soil moisture across the 2016–2024 period, potentially reflecting gradual hydroclimatic shifts, though its contribution remains secondary to spatial and climatic drivers. This discordance between linear and nonlinear methods demonstrates that soil moisture dynamics are governed by complex relationships that cannot be fully captured by Pearson correlation alone.

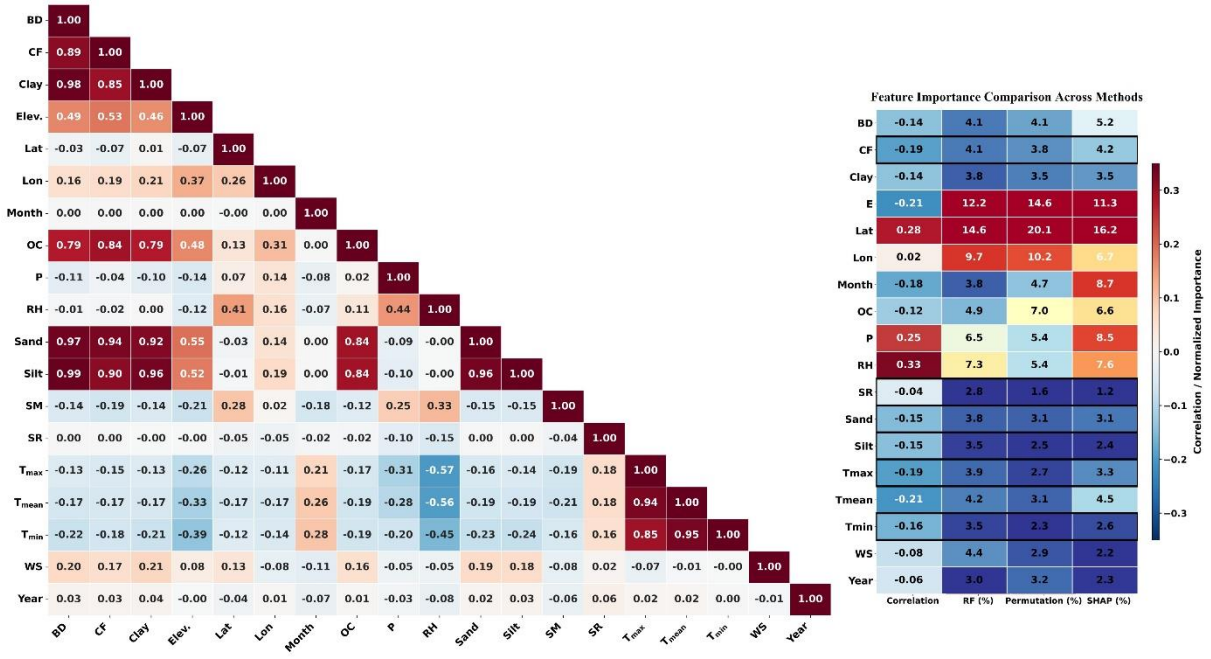


Figure 3- Multi-method feature selection analysis at 20 cm depth: (left) Pearson correlation matrix among predictors and soil moisture, and (right) comparative feature importance rankings from Pearson correlation, Random Forest, Permutation Importance, and SHAP analyses

Multicollinearity was rigorously addressed through a combination of correlation thresholds (|r| > 0.80) and VIF analysis. Six variables were excluded due to redundancy: Maximum and Minimum Temperature (both r ≈ 0.94 with Mean Temperature), Silt (r = 0.99 with Bulk Density), Sand (r = 0.97 with Bulk Density), and Coarse Fragments (r = 0.89 with Bulk Density). VIF analysis further confirmed severe multicollinearity for these predictors, providing additional justification for their removal. SR was also excluded due to its weak linear relationship with soil moisture (r = −0.04) and consistently low importance across all methods (RF: 2.8%, Permutation: 1.6%, SHAP: 1.2%).

Crucially, although Latitude and Longitude exhibited high Variance Inflation Factors (VIF > 100), they were retained. This multicollinearity arises from geographic dependency inherent to Türkiye's climate zones, where Latitude correlates with precipitation gradients and Longitude with continentality. Despite high VIF values, these variables were retained because they capture distinct climatic processes essential for soil moisture prediction, as confirmed by their consistently high importance across Random Forest, Permutation, and SHAP analyses.

Year also exhibited high VIF (>914), though it showed weak individual correlations with other predictors. This elevated VIF likely reflects the combined predictability of Year from multiple spatiotemporal variables rather than true redundancy. Year was retained despite modest importance (SHAP:2.3%, RF:3.0%) because it captures interannual climate variability and potential long-term hydroclimatic trends over the 2016–2024 period that cannot be represented by spatial or monthly predictors alone.

Similarly, Bulk Density showed high VIF (1069) due to its mathematical relationship with particle size distribution but was retained as it directly represents soil physical structure—a fundamental control on water retention capacity. Among soil texture components, Clay was retained (RF: 3.8%, Permutation: 3.5%, SHAP: 3.5%) as it represents the primary texture control on water retention following the removal of Sand and Silt, despite its modest linear correlation (r = −0.14). The final predictor set consisted of 12 variables. These 12 variables are distributed across three functional categories, with importance values averaged across the three feature importance methods (RF, Permutation, SHAP). Spatiotemporal variables (Year, Latitude, Longitude, Elevation, Month) accounted for approximately 47% of combined importance (43.3%, 52.8%, 45.2% for the three methods, respectively), climatic variables (Relative Humidity, Precipitation, Mean Temperature, Wind Speed) contributed around 21% (22.3%, 16.8%, 22.8%), and soil properties (Bulk Density, Clay, Organic Carbon) represented about 13% (12.3%, 13.6%, 13.2%). This distribution demonstrates that soil moisture prediction requires integrating spatiotemporal gradients, climatic drivers, and soil properties, reflecting the multi-scale nature of soil moisture controls.

The substantial importance of geographic variables reflects Türkiye's pronounced spatial heterogeneity. Latitude, which ranked among the top predictors across all tree-based methods, captures the north–south climatic gradient, ranging from humid Black Sea conditions to the semi-arid southeastern plains. Longitude reflects the west–east continentality gradient, while Elevation controls orographic precipitation enhancement and seasonal snow accumulation. These spatial variables collectively demonstrate that Türkiye encompasses multiple distinct hydrological zones that fundamentally shape soil moisture patterns, despite its modest latitudinal span.

The combined feature selection procedure successfully captured both linear relationships and nonlinear interactions essential for soil moisture prediction. By integrating Pearson correlation, VIF-based multicollinearity screening, and three independent feature-importance methods (RF, Permutation, SHAP), the analysis identified genuinely informative predictors while eliminating redundant variables. This multi-method approach addresses potential biases in individual methods (e.g., RF's impurity-based preference for high-cardinality variables) and ensures more reliable rankings. The consistent prominence of key features—particularly Latitude, Elevation, RH, and Precipitation—provides strong evidence of their relevance, although feature rankings may vary across different modeling approaches.

3.3. Machine learning model comparison

3.3.1. Algorithm performance comparison

Performance evaluation of seven machine learning algorithms revealed substantial variation in predictive accuracy, with R^2 values ranging from 0.27 to 0.74 (Figure 4). This wide performance spectrum indicates that different algorithms vary considerably in their ability to capture the complex, nonlinear interactions governing soil moisture dynamics. The results are consistent with previous findings (Houben et al. 2025), confirming the superiority of ensemble methods over conventional approaches for soil moisture prediction.

LR and SVR yielded the weakest performance (R^2 = 0.29 and 0.27, respectively). Linear Regression's poor performance (NSE = 0.29, KGE = 0.33) is expected given the inherently nonlinear nature of soil moisture dynamics. SVR also underperformed (NSE = 0.27, KGE = 0.28, Willmott d = 0.62) despite its theoretical capacity for nonlinear modeling, consistent with previous hydrological studies reporting lower SVR accuracy relative to ensemble methods (Hao & Bai 2023).

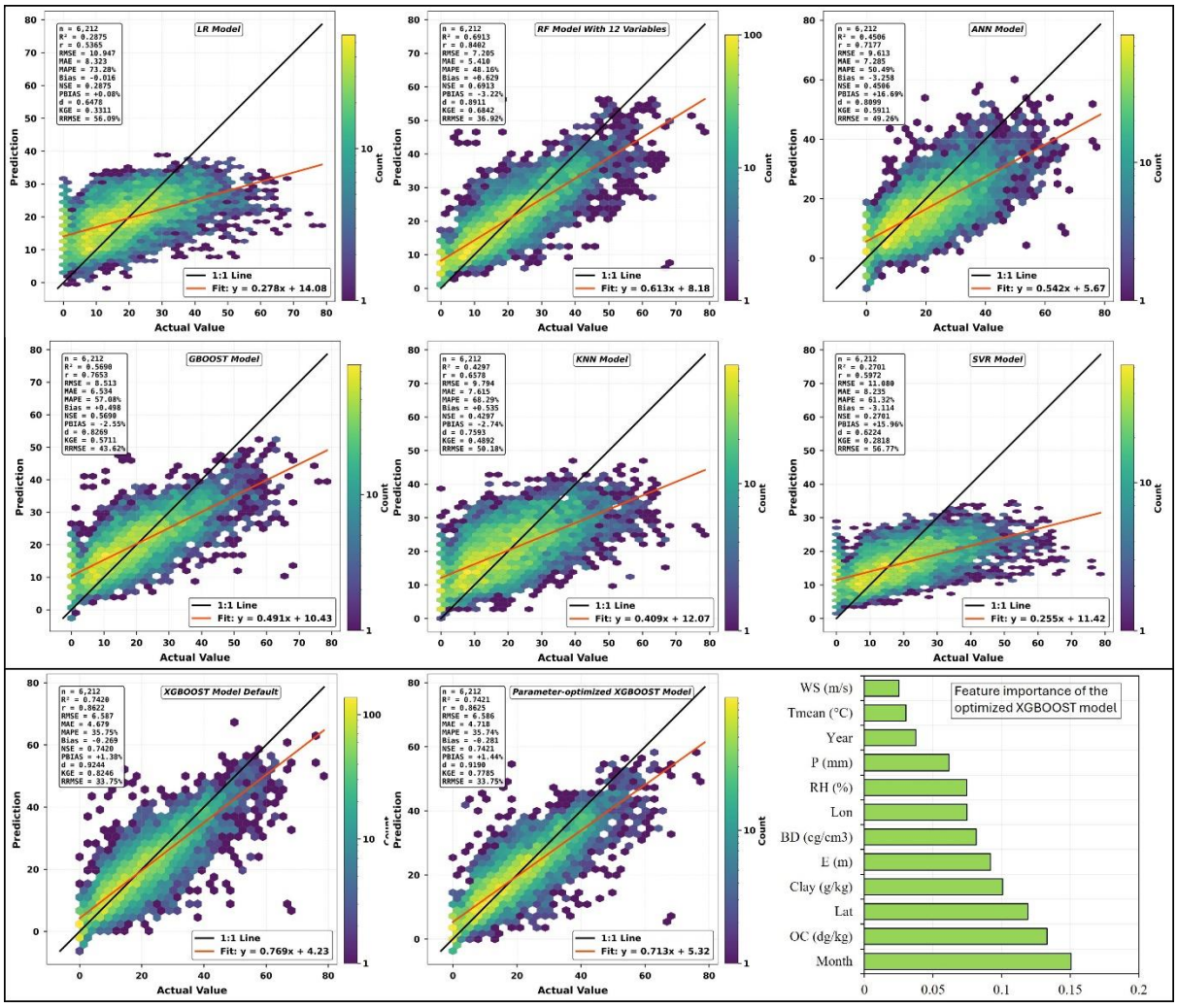


Figure 4- Performance comparison of seven machine learning algorithms at 20 cm and feature importance rankings from the optimized XGBoost model (bottom panel: percentage contribution of 12 predictors)

ANN and KNN achieved intermediate accuracy (R^2 = 0.45 and 0.43, respectively). ANN's limited performance (PBIAS= %16.7, RMSE = 9.61%) may reflect insufficient training epochs or suboptimal architecture for this dataset. KNN's limitation arises from

its reliance on local similarity. As also noted by Zeraatpisheh et al. (2025), this approach performs poorly in spatially and temporally heterogeneous systems, where nearby observations may exhibit dissimilar soil moisture dynamics. GB outperformed both ANN and KNN through its ensemble approach (R^2=0.57), though its default parameter configuration still limited further improvement. These intermediate results demonstrate the progressive advantage of ensemble methods over instance-based or single-network approaches.

The ensemble-based models, RF and XGBoost, achieved the highest performance with default configurations (R^2 = 0.69 and 0.74, respectively). RF demonstrated strong performance across complementary metrics (NSE = 0.69, KGE = 0.68, Willmott d = 0.89), while XGBoost showed superior performance (NSE = 0.74, KGE = 0.82, Willmott d = 0.92). PBIAS remained within acceptable limits for both models (RF: -3.2%, XGBoost: 1.4%), indicating minimal systematic over- or under-prediction. However, both exhibited overfitting concerns: RF showed a train–test R^2 difference (ΔR^2) of 0.16, while XGBoost displayed more pronounced overfitting with ΔR^2 = 0.20. These results indicate that although decision-tree ensembles effectively capture nonlinear soil moisture dynamics, their default regularization settings are insufficient for optimal generalization.

Hyperparameter optimization revealed contrasting sensitivities. RF showed negligible improvement ($\Delta R^2 < 0.01$), retaining $R^2 \approx 0.69$, consistent with its inherent robustness from bootstrap aggregation. However, overfitting persisted (ΔR^2 = 0.16), suggesting that RF's structural design limits further complexity regulation through hyperparameters. In contrast, XGBoost benefited substantially—not in predictive accuracy, but in generalization capability. Using RandomizedSearchCV with 100 parameter combinations and 5-fold temporal cross-validation (TimeSeriesSplit), the optimized model maintained test performance (R^2 = 0.74, RMSE = 6.59%, MAE = 4.72%) while reducing overfitting from ΔR^2 = 0.20 to 0.13. The optimized model's hydrology-standard metrics (NSE = 0.74, KGE = 0.78, Willmott d = 0.92, PBIAS = 1.4%, RRMSE = 33.7%) confirmed satisfactory performance. Detailed hyperparameter configurations for all models are provided in Supplementary Table S1.

Strict temporal validation protocols were maintained throughout model development to prevent data leakage. The dataset was split chronologically (2016–2021 training, 2022–2024 testing) without overlap. All preprocessing, standardization, and hyperparameter optimization using 5-fold TimeSeriesSplit were performed exclusively on training data.

At 20 cm depth, the optimized XGBoost model achieved cross-validation R^2 of 0.65 and test R^2 of 0.74. This pattern reflects the structure of TimeSeriesSplit combined with temporal data availability. Early folds train on limited observations from 2016–2017, when the monitoring network was still expanding and many stations had incomplete records. These data-scarce early folds produce conservative performance estimates. Later folds progressively incorporate more complete datasets from 2018–2021, approaching full training performance. The final model, trained on the complete 2016–2021 period, achieved training R^2 of 0.87. The training-test difference of 0.13 confirms adequate model complexity without overfitting.

3.3.2. Feature importance analysis

Feature importance patterns from the two best-performing algorithms were systematically compared using the final 12-variable configuration established in Section 3.1 (Figure 3 and 4). RF assigned the highest importance to Latitude (19.1%), Elevation (14.4%), and Longitude (12.2%), highlighting spatial positioning as the primary driver of surface soil moisture variability. In contrast, the optimized XGBoost model prioritized Seasonality (Month, 15.1%), Organic Carbon (13.3%), Latitude (11.9%), and Clay content (10.1%), reflecting enhanced sensitivity to soil physical properties and temporal dynamics.

XGBoost assigned greater importance to soil properties compared to RF. Organic Carbon increased from 7.9% (RF) to 13.3% (XGBoost), Clay from 6.3% to 10.1%, and Bulk Density from 6.0% to 8.2%. This elevation in ranking reflects XGBoost's enhanced capacity to resolve complex, threshold-based interactions in soil hydraulic properties. Although high-cardinality variables can theoretically inflate importance scores by offering more split points (Zhang, 2023), the high ranking of these soil properties is firmly supported by hydrological principles. Specifically, Organic Carbon, Clay, and Bulk Density directly control field capacity, wilting point, and hydraulic conductivity, which are fundamental determinants of soil water storage capacity. The physical plausibility of these rankings, combined with their consistent identification across multiple feature selection methods in Section 3.1 (Permutation Importance and SHAP analyses), confirms their genuine hydrological relevance rather than algorithmic artifact.

In contrast, RF gave greater weight to spatiotemporal predictors (Latitude (19.1%), Elevation (14.4%), and Longitude (12.2%)) which collectively represented 45.6% of total importance in RF versus 28.6% in XGBoost. These variables capture broad-scale climatic gradients and topographic controls across Türkiye's heterogeneous landscape. Despite the methodological differences, both models consistently ranked Latitude and Elevation among the top predictors, with combined importance of 33.5% (RF) and 21.1% (XGBoost). This convergence was independently corroborated by Permutation Importance (Latitude: 20.1%, Elevation: 12.8%) and SHAP analysis (Latitude: 16.2%, Elevation: 11.5%), which confirmed these spatial variables as dominant controls through model-agnostic evaluation. The multi-method agreement provides robust evidence that Türkiye's pronounced north-to-south climatic gradient—ranging from humid coastal regions (>2000 mm annual precipitation) to semi-arid interior areas (<300 mm)—and extreme topographic variability (0 to 5137 m elevation range) exert primary control over regional soil moisture distribution.

These differences arise from fundamental distinctions in how each algorithm calculates feature importance. RF evaluates each feature's average contribution to predictive accuracy across trees based on impurity reduction, whereas XGBoost computes the reduction in total model loss attributable to each feature through gradient-based optimization (Zhang, 2023). XGBoost's formulation enables more precise detection of nonlinear interactions and threshold-based effects through its iterative error correction mechanism and regularization framework (Xia et al. 2024).

On the other hand, climatic variables (Precipitation, Relative Humidity, Temperature and Wind Speed) showed moderate importance in both models (RF: 24.8%; XGBoost: 19.2%). This relatively modest contribution likely results from the use of monthly averages, which smooth out short-term variability and extreme events that drive rapid changes in soil moisture. Higher-resolution (daily or weekly) data would better capture these short-term dynamics and likely strengthen these variables' predictive power. However, monthly data enhances model applicability in operational water management, given their broader availability and lower computational demand.

3.4. Temporal and spatial analysis

The optimized XGBoost model was applied to assess soil moisture dynamics across seasonal and regional subgroups to evaluate performance stability and feature importance under varying environmental conditions. The analysis revealed systematic variations in predictive accuracy and feature importance rankings (Figure 5).

3.4.1. Seasonal variability

Model performance exhibited clear seasonal patterns, with R^2 values ranging from 0.60 to 0.69. The highest predictive accuracy was obtained in winter (R^2 = 0.69) and spring (R^2 = 0.68), whereas lower accuracy was observed in summer and autumn (R^2 = 0.61 and 0.60, respectively). This performance gradient reflects differences in the underlying stability of soil moisture dynamics. Summer and autumn are characterized by high temperatures and intense evapotranspiration combined with irregular precipitation, producing rapid fluctuations in soil moisture that are difficult to capture using monthly averaged predictors. In contrast, winter and spring experience more consistent precipitation patterns and lower evaporative demand, resulting in smoother temporal variations that enhance model predictive reliability.

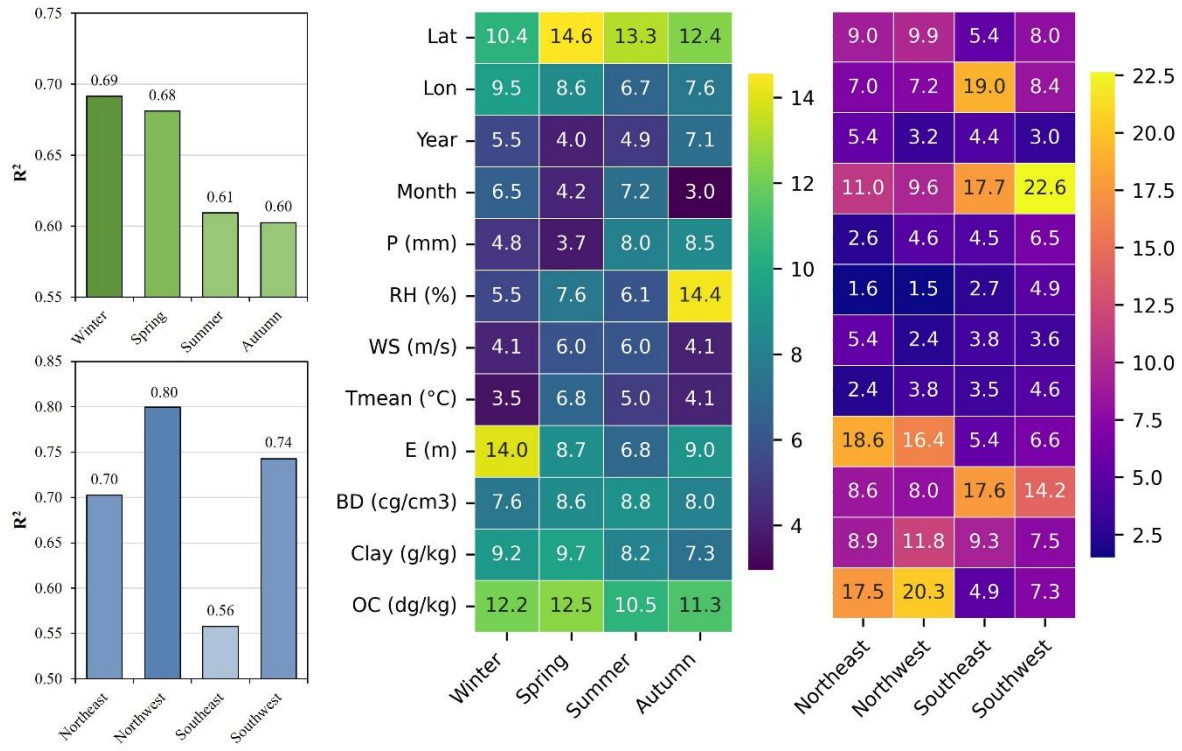


Figure 5- Seasonal and regional variability in XGBoost model performance and feature importance at 20 cm depth: test R^2 values by season (top-left) and region (bottom-left), and heatmaps showing relative importance (%) of 12 predictors across four seasons (center) and four geographic regions of Türkiye (right)

Feature importance rankings also varied markedly between seasons, revealing systematic shifts in the dominant controls on soil moisture (Figure 5). Elevation dominated during winter (14%), reflecting its control over snow accumulation, melt timing, and orographic precipitation enhancement. Latitude emerged as the most influential predictor during spring (14.6%) and summer

(13.3%), consistent with the intensified role of the north–south climatic gradient under increasing evapotranspiration demand. During autumn, Relative Humidity exhibited the highest importance (14.4%), indicating heightened sensitivity to atmospheric moisture conditions.

Despite these seasonal shifts in the dominant predictor, Organic Carbon maintained consistently high importance across all seasons (10.5–12.5%), confirming the fundamental role of soil water retention capacity in regulating moisture dynamics throughout the year. Latitude also showed sustained importance (10.4–14.6%), underscoring that Türkiye's pronounced north–south climatic gradient exerts a persistent structural control on soil moisture regardless of season.

The collective importance of climatic variables (precipitation, relative humidity, wind speed, and temperature) exhibited a progressive increase throughout the year, rising from 17.9% in winter to 24.1% in spring, 25.1% in summer, and peaking at 31.1% in autumn. This pattern indicates heightened sensitivity of soil moisture to meteorological fluctuations during the dry-down and subsequent rewetting phases. Specifically, the peak importance in autumn reflects the critical role of atmospheric conditions when the soil moisture deficit accumulated during summer makes the system highly responsive to precipitation and humidity.

3.4.2. Spatial variability

Regional model performance ranged from $R^2 = 0.56$ to 0.80 (Figure 5). The Northwest exhibited the highest accuracy ($R^2 = 0.80$), followed by the Southwest ($R^2 = 0.74$) and Northeast ($R^2 = 0.70$), while the Southeast showed the lowest performance ($R^2 = 0.56$). This variation primarily reflects differences in station density, climatic heterogeneity, and the inherent predictability of soil moisture dynamics across distinct climate zones. The Southeast's semi-arid climate, with irregular rainfall and extreme temperatures, imposes greater challenges for prediction compared to the more stable and humid Northwest.

Feature importance patterns varied substantially among regions (Figure 5). Organic Carbon exhibited the highest importance in the Northwest (20.3%), decreasing markedly in other regions (Northeast: 17.5%, Southwest: 7.3%, Southeast: 4.9%). The dominance of organic carbon as a predictor in the Northwest reflects the region's humid climate and associated high organic matter content. In the Northeast, Elevation was the most influential variable (18.6%), consistent with the mountainous terrain of Eastern Anatolia, where orography governs precipitation distribution and snowmelt timing. In contrast, Elevation showed lower importance in the Southwest (6.6%) and Southeast (5.4%).

Month (seasonality) exhibited pronounced regional variation, ranging from 9.6% in the Northwest to 22.6% in the Southwest. In the Southwest's Mediterranean climate with distinct seasonal patterns, temporal dynamics become particularly important for capturing moisture variability. Similarly, in the Southeast (17.7%), seasonality reflects the strong contrast between hot, dry summers and cooler, wetter winters characteristic of semi-arid continental climates.

Soil properties collectively (Organic Carbon, Bulk Density, Clay) exhibited a clear spatial gradient, contributing 40.1% in the Northwest, 35.0% in the Northeast, 29.0% in the Southwest, and 31.8% in the Southeast. This pattern indicates that soil physical characteristics exert stronger control on moisture dynamics in humid regions with abundant organic matter, whereas in semi-arid areas, the relative influence of climatic drivers increases significantly, indicating that water availability becomes highly sensitive to instantaneous meteorological conditions.

Climatic variables collectively contributed between 11.9% and 19.5%, with higher importance observed in the Southwest (19.5%) and Southeast (14.5%). In the Southeast, this elevated importance reflects the region's semi-arid climate, characterized by limited and irregular precipitation combined with extreme temperatures. In the Southwest, the Mediterranean climate with distinct dry summers increases dependence on precipitation and atmospheric humidity for maintaining adequate soil moisture.
Overall, the temporal and spatial analyses demonstrate that the controlling factors of soil moisture vary systematically across seasons and regions. Despite this heterogeneity, model performance remained consistently high across seasons ($R^2 = 0.60–0.69$) and showed strong performance in three of four regions ($R^2 = 0.70–0.80$), with lower accuracy only in the Southeast ($R^2 = 0.56$). This confirms XGBoost's robustness and adaptability across diverse environmental conditions. The observed variations in feature importance reveal how different climatic, topographic, and edaphic processes interact to shape soil moisture dynamics across Türkiye's heterogeneous landscapes.

3.5. Depth-dependent variations in soil moisture prediction and feature importance

Following the demonstration of XGBoost's superior performance at the 20 cm depth, the same algorithm was applied to develop independent models for the 40 cm and 80 cm layers. This extension enables assessment of how soil moisture prediction accuracy and controlling factors vary across the vertical profile. However, data availability at these deeper layers was substantially more limited (n = 2,280 for each depth compared to n = 17,905 at 20 cm), reflecting the operational structure of the monitoring network where deep soil moisture measurements are conducted at fewer stations. This limitation precluded the seasonal and regional sub-group analyses conducted for the surface layer.

Model performance exhibited a progressive decline across depths (Figure 6). Test R^2 decreased from 0.74 at 20 cm to 0.69 at 40 cm and 0.66 at 80 cm. This vertical gradient in performance reflects fundamental hydrological processes. In deeper soil layers, moisture transport slows considerably, spatial heterogeneity increases, and direct coupling with surface meteorological conditions weakens (Zhang et al. 2025). Despite this reduction, all three depths maintained satisfactory predictive performance ($R^2 > 0.65$), demonstrating reasonable capability for capturing root-zone moisture dynamics relevant to agricultural water management applications.

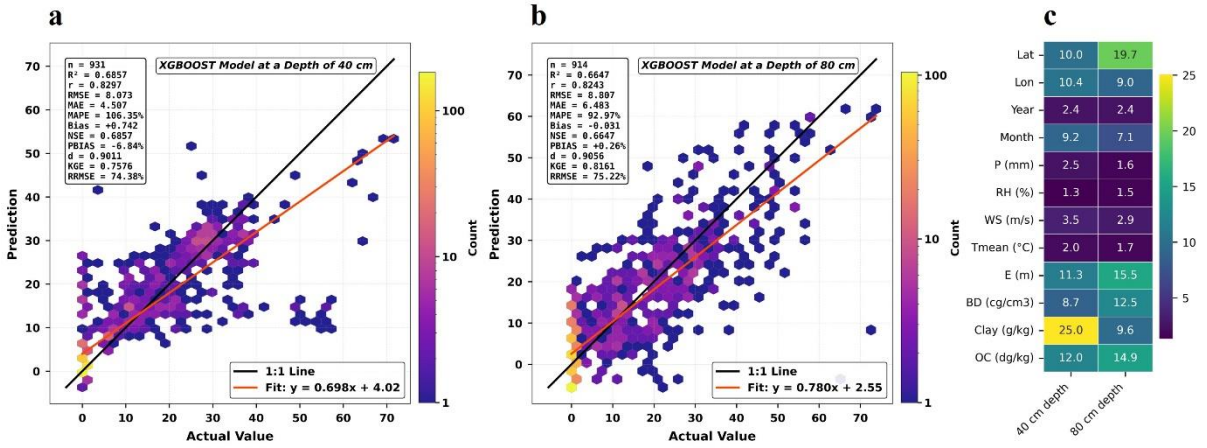


Figure 6- Multi-depth model performance and feature importance: scatter plots of predicted vs observed soil moisture for (a) 40 cm and (b) 80 cm depths, with performance metrics (R^2, RMSE, NSE, KGE) and (c) heatmap showing feature importance rankings across two depths (40, and 80 cm)

The relative importance of predictors shifted markedly across depths. At 20 cm, Month (15.1%) and Organic Carbon (13.3%) were the most important predictors, reflecting the dominant roles of seasonal dynamics and surface soil properties in controlling near-surface moisture. At 40 cm, Clay content emerged as the most influential variable (25.0%), followed by Organic Carbon (12.0%) and Elevation (11.3%). The dominance of clay at this intermediate depth indicates that soil texture becomes the primary control on water retention capacity within the root-zone transition layer. At 80 cm, Latitude became the most important predictor (19.7%), followed by Elevation (15.5%) and Organic Carbon (14.9%). This shift reflects a fundamental transition: surface moisture responds primarily to seasonal forcing and local soil properties, whereas deeper layers integrate longer-term regional climate patterns represented by the north–south gradient and topographic controls across Türkiye.

Organic Carbon (OC) exhibited a U-shaped pattern across depths: 13.3% at 20 cm, 12.0% at 40 cm, and 14.9% at 80 cm. At the surface (20 cm), OC directly enhances water retention by improving soil structure and porosity. At 40 cm, clay becomes the dominant control on moisture storage (25.0%), reducing OC's relative influence. At 80 cm, OC importance rises again (14.9%), but for a different reason: it functions as a spatial indicator of long-term climate rather than a physical water retainer. Surface OC content (0–5 cm) varies systematically across Türkiye's climatic gradient, with high values in humid northern regions (>2.5%) and low values in semi-arid southern areas (<1.2%). Deep soil moisture (80 cm) exhibits a similar spatial distribution. At this depth, where short-term weather fluctuations are attenuated, surface OC effectively captures this regional climatic zonation and serves as a spatial proxy alongside Latitude and Elevation.

Climatic variables (precipitation, relative humidity, temperature, wind speed) showed a pronounced decrease with depth. Their collective importance declined from 19.1% at 20 cm to 9.2% at 40 cm and 7.8% at 80 cm. This pattern indicates that deeper soil moisture becomes progressively insulated from short-term atmospheric forcing. Surface layers (20 cm) respond directly to daily and weekly weather events. Deeper layers reflect long-term climatic regimes rather than individual meteorological fluctuations. These depth-dependent patterns reveal a fundamental vertical stratification in moisture controls. At the surface (20 cm), soil moisture responds rapidly to short-term weather events and is strongly influenced by local soil properties like organic carbon. At intermediate depth (40 cm), clay content dominates moisture storage. At the deepest layer (80 cm), moisture dynamics are governed primarily by long-term climatic regimes represented by geographic position. This vertical stratification has important implications for prediction strategies and agricultural water management. Surface layers require high-temporal-resolution weather data, whereas deeper root-zone layers can be reliably modeled using climatological averages and geographic descriptors. Accurate prediction of soil moisture across multiple depths provides essential information for agricultural water management. Quantifying total root-zone water storage enables optimization of irrigation scheduling, early detection of drought stress, and prevention of yield losses in water-sensitive crops. The multi-layer modeling framework developed in this study allows comprehensive characterization of soil moisture throughout the profile, providing an operational foundation for decision support systems aimed at enhancing water-use efficiency and supporting sustainable agricultural production.

4. Conclusions

This study demonstrates that machine learning offers a robust and scalable solution for predicting soil moisture across multiple depths using widely accessible environmental data. Among the seven algorithms tested, XGBoost consistently outperformed alternatives, achieving high accuracy at 20, 40, and 80 cm depths (R^2 = 0.74, 0.69, and 0.66, respectively). Beyond overall performance, the depth-resolved analysis revealed a clear shift in the dominant controls of soil moisture: surface layers were primarily influenced by seasonal variations and local soil properties (particularly Month and Organic Carbon) while subsurface moisture was increasingly governed by large-scale geographic and climatic gradients, such as latitude and elevation. This transition from short-term, event-driven dynamics at the surface to long-term, regime-driven processes at depth reflects fundamental hydrological stratification with direct implications for soil moisture monitoring and water management strategies.
The proposed framework addresses longstanding limitations of conventional soil moisture measurement methods, which are often constrained by cost, spatial sparsity, and logistical complexity. The model utilizes readily available predictors that are either routinely monitored (meteorological variables), already mapped through remote sensing products (soil properties such as organic carbon, bulk density, and clay content), or inherently known (geographic coordinates and elevation). This enables generation of spatially continuous and depth-specific moisture estimates without requiring soil moisture sensors at every prediction location. The approach is particularly valuable in data-limited regions where establishing dense soil moisture monitoring networks is prohibitively expensive, yet root-zone information remains critical for effective decision-making. The model supports practical applications including precision irrigation scheduling, drought risk assessment, and regional water resource planning. Its strong and stable performance across Türkiye's varied hydro-climatic zones (from humid coastal regions to semi-arid interior areas) demonstrates robustness under heterogeneous conditions and highlights potential for adaptation in other water-stressed agricultural environments.

Several methodological and data-related constraints must be acknowledged. The reliance on monthly-averaged meteorological inputs represents a significant temporal limitation. Although suitable for capturing seasonal patterns, such coarse resolution cannot fully reflect rapid hydrological responses such as storm-driven infiltration or short-lived evapotranspiration spikes. This reduces model sensitivity to extreme events and short-term moisture fluctuations. Additionally, the pronounced imbalance in observational data across soil depths poses challenges for subsurface predictions. Sample sizes at 40 and 80 cm are approximately 85% lower than at 20 cm, which increases predictive uncertainty and limits the robustness of depth-specific generalizations. This constraint is particularly critical given the importance of root-zone moisture for agricultural applications. Furthermore, the current framework has not been evaluated under non-stationary climate conditions or in regions experiencing systematic land-use changes. Model performance under such evolving conditions remains uncertain and requires careful validation before operational deployment.

Future research should prioritize three key directions to enhance practical applicability. First, incorporating higher temporal resolution meteorological data (daily or sub-daily) combined with satellite-based observations from missions such as SMAP and Sentinel-1 would improve model sensitivity to rapid moisture dynamics. This integration could also address spatial data gaps in regions with sparse ground monitoring networks. Second, the framework should be evaluated under future climate scenarios using downscaled projections from Global Climate Models. This would assess potential shifts in soil moisture regimes under altered temperature and precipitation patterns. Such analysis is essential for supporting climate-resilient agricultural planning and water resource management. Third, operational validation through real-time implementation in irrigation scheduling systems or drought early-warning platforms would provide critical feedback on practical performance. Field-based testing would identify necessary refinements and establish confidence intervals appropriate for decision-support applications.

This study establishes a reliable and interpretable machine learning framework for multi-depth soil moisture prediction in complex environments. By quantifying how environmental drivers vary with depth, the framework advances understanding of soil moisture dynamics and provides a practical tool for sustainable agricultural water management. The demonstrated success across Türkiye's diverse landscapes offers strong potential for broader global application, particularly in regions facing increasing water scarcity and climate-induced hydrological variability.

Nomenclature

ANN	:	Artificial Neural Network	P	:	Monthly Total Precipitation (mm)
BD	:	Bulk Density (cg/cm^3)	PBIAS	:	Percent Bias (%)
Clay	:	Clay Content (g/kg)	R^2	:	Coefficient of Determination
CV	:	Cross-Validation	RF	:	Random Forest
DT	:	Decision Tree	RH	:	Relative Humidity (%)
Elev	:	Elevation (m)	RMSE	:	Root Mean Square Error (%)
KNN	:	K-Nearest Neighbors	RRMSE	:	Relative Root Mean Square Error (%)
Lat	:	Latitude (°)	SHAP	:	SHapley Additive exPlanations
Lon	:	Longitude (°)	SM	:	Soil Moisture (%)
LR	:	Linear Regression	SVM	:	Support Vector Machine
MAE	:	Mean Absolute Error (%)	T_{mean}	:	Mean Air Temperature (°C)
ML	:	Machine Learning	VIF	:	Variance Inflation Factor
NSE	:	Nash-Sutcliffe Efficiency	WS	:	Wind Speed (m/s)
OC	:	Organic Carbon Content (dg/kg)	XGBoost	:	Extreme Gradient Boosting

Author Declarations

Authors' contributions: The author (M.S.D.) contributed to all aspects of the study, including conceptualization, methodology, data collection, analysis, and manuscript writing.

Conflict of interest/Competing interests: The author declares no conflict of interest.

Ethics approval/declarations: No need/ Not applicable.

Funding: This research received no external funding.

Consent to participate: No need/ Not applicable.

Consent for publication: No need/ Not applicable.

Data availability: Soil moisture and meteorological data were obtained from the Turkish State Meteorological Service (Meteoroloji Genel Müdürlüğü, MGM). Soil property data were obtained from the SoilGrids250m global dataset (https://soilgrids.org). Model hyperparameters and configurations are documented in Supplementary Tables S1 and S2. Additional data supporting the findings are available from the author upon reasonable request.

References

Ahmad S, Kalra A & Stephen H (2010). Estimating soil moisture using remote sensing data: A machine learning approach. *Advances in Water Resources* 33(1): 69–80. https://doi.org/10.1016/j.advwatres.2009.10.008

Ali S, Khorrami B, Jehanzaib M, Tariq A, Ajmal M, Arshad A, Shafeeque M, Dilawar A, Basit I, Zhang L, Sadri S, Niaz M A, Jamil A & Khan S N (2023). Spatial Downscaling of GRACE Data Based on XGBoost Model for Improved Understanding of Hydrological Droughts in the Indus Basin Irrigation System (IBIS). *Remote Sensing* 15(4): 873. https://doi.org/10.3390/rs15040873

Allen R, Pereira L, Raes D & Smith M (1998). FAO Irrigation and Drainage Paper No. 56. Bayazıt M (2013). *Hidroloji (1st ed.)*. Istanbul: Birsen Yayınevi.

Cengiz A I, Güven H, Erşahin M E, Özgün H, Tuna M & Kinaci C (2022). Investigation of Factors Affecting Per Capita Expenditure of Water and Sewerage Administrations in Turkey. *Turkish Journal of Water Science and Management*. https://doi.org/10.31807/tjwsm.1182391

Champagne C, White J, Berg A, Belair S & Carrera M (2019). Impact of Soil Moisture Data Characteristics on the Sensitivity to Crop Yields Under Drought and Excess Moisture Conditions. *Remote Sensing* 11(4): 372. https://doi.org/10.3390/rs11040372

Chen L, Hu B, Sun J, Xu Y J, Zhang G, Ma H & Ren J (2025). Using remote sensing and machine learning to generate 100-cm soil moisture at 30-m resolution for the black soil region of China: Implication for agricultural water management. *Agricultural Water Management 309*: 109353. https://doi.org/10.1016/j.agwat.2025.109353

Chen T & Guestrin C (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Guo J, Zhang F, Li W, Yang A, Fan Y & Li J (2025). Runoff Prediction in the Xiangxi River Basin Under Climate Change: The Application of the HBV-XGBoost Coupled Model. *Water 17*(16): 2420. https://doi.org/10.3390/w17162420

Hao R & Bai Z (2023). Comparative Study for Daily Streamflow Simulation with Different Machine Learning Methods. *Water* 15(6): 1179. https://doi.org/10.3390/w15061179

Harani P, Gautam S, Joshi S K, Asirvatham L G & Rakshith B L (2025). Advancements in soil moisture estimation through integration of remote sensing and artificial intelligence techniques. *Science of The Total Environment* 1001: 180503. https://doi.org/10.1016/j.scitotenv.2025.180503

Houben T, Ebeling P, Khurana S, Schmid J S & Boog J (2025). Machine-learning based spatiotemporal prediction of soil moisture in a grassland hillslope. *Vadose Zone Journal* 24(2): e70011. https://doi.org/10.1002/vzj2.70011

Irmak S (2014). Plant Growth and Yield as Affected by Wet Soil Conditions Due to Flooding or Over-Irrigation. University of Nebraska—Lincoln Extension. http://www.ianrpubs.unl.edu

Karthikeyan L & Mishra A K (2021). Multi-layer high-resolution soil moisture estimation using machine learning over the United States. *Remote Sensing of Environment* 266: 112706. https://doi.org/10.1016/j.rse.2021.112706

Li C, Zhou Y, Wu M, Xu J & Fu X (2025). Exploring Nonlinear Threshold Effects and Interactions Between Built Environment and Urban Vitality at the Block Level Using Machine Learning. *Land* 14(6): 1232. https://doi.org/10.3390/land14061232

Ma M, Zhao G, He B, Li Q, Dong H, Wang S & Wang Z (2021). XGBoost-based method for flash flood risk assessment. *Journal of Hydrology* 598: 126382. https://doi.org/10.1016/j.jhydrol.2021.126382

Mdemu M, Kissoly L, Bjornlund H, Kimaro E, Christen E W, Van Rooyen A, Stirzaker R & Ramshaw P (2020). The role of soil water monitoring tools and agricultural innovation platforms in improving food security and income of farmers in smallholder irrigation schemes in Tanzania. *International Journal of Water Resources Development* 36(sup1): S148–S170. https://doi.org/10.1080/07900627.2020.1765746

Meteoroloji Genel Müdürlüğü. (MGM) (2023). Areal precipitation assessment for the 2023 water year (In Turkish). Republic of Türkiye Ministry of Environment, Urbanization and Climate Change, Research Department, Hydrometeorology Division.

Meteoroloji Genel Müdürlüğü. (MGM) (2024). Average temperature in Türkiye: Analysis for the period 1970–2024 (In Turkish). Republic of Türkiye Ministry of Environment, Urbanization and Climate Change.

Nguyen T T, Ngo H H, Guo W, Chang S W, Nguyen D D, Nguyen C T, Zhang J, Liang S, Bui X T & Hoang N B (2022). A low-cost approach for soil moisture prediction using multi-sensor data and machine learning algorithm. *Science of The Total Environment* 833: 155066. https://doi.org/10.1016/j.scitotenv.2022.155066

Nichols S (2011). Review and evaluation of remote sensing methods for soil-moisture estimation. *Journal of Photonics for Energy*: 028001. https://doi.org/10.1117/1.3534910

Ölgen M K (2010). Spatial distribution of annual and seasonal precipitation variability in Türkiye (In Turkish). Ege Coğrafya Dergisi 19(1): 85-95.

Rasheed M W, Tang J, Sarwar A, Shah S, Saddique N, Khan M U, Imran Khan M, Nawaz S, Shamshiri R R, Aziz M & Sultan M (2022). Soil Moisture Measuring Techniques and Factors Affecting the Moisture Dynamics: A Comprehensive Review. *Sustainability* 14(18): 11538. https://doi.org/10.3390/su141811538

Romero Martínez M, Carmona Ibáñez P & Martínez Vargas J (2025). Predicting Business Failure with the XGBoost Algorithm: The Role of Environmental Risk. *Sustainability* 17(11): 4948. https://doi.org/10.3390/su17114948

Seleiman M F, Al-Suhaibani N, Ali N, Akmal M, Alotaibi M, Refay Y, Dindaroglu T, Abdul-Wajid H H & Battaglia M L (2021). Drought Stress Impacts on Plants and Different Approaches to Alleviate Its Adverse Effects. *Plants* 10(2): 259. https://doi.org/10.3390/plants10020259

Özütürk S (2024). Plants under the risk drought and desertification in southeastern Anatolia (Turkey). *EPRA International Journal of Multidisciplinary Research (IJMR)*: 533–544. https://doi.org/10.36713/epra18002

Susha Lekshmi S U, Singh, D N & Shojaei Baghini, M. (2014). A critical review of soil moisture measurement. *Measurement* 54: 92–105. https://doi.org/10.1016/j.measurement.2014.04.007

Sun W, Zhou S, Yu B, Zhang Y, Keenan T & Fu B (2025). Soil moisture-atmosphere interactions drive terrestrial carbon-water trade-offs. *Communications Earth & Environment* 6(1): 169. https://doi.org/10.1038/s43247-025-02145-z

Szczepanek R (2022). Daily Streamflow Forecasting in Mountainous Catchment Using XGBoost, LightGBM and CatBoost. *Hydrology 9*(12): 226. https://doi.org/10.3390/hydrology9120226

Taktikou E, Bourazanis G, Papaioannou G & Kerkides P (2016). Prediction of Soil Moisture from Remote Sensing Data. *Procedia Engineering 162*: 309–316. https://doi.org/10.1016/j.proeng.2016.11.066

Teshome F T, Bayabil H K, Schaffer B, Ampatzidis Y & Hoogenboom G (2024). Improving soil moisture prediction with deep learning and machine learning models. *Computers and Electronics in Agriculture* 226: 109414. https://doi.org/10.1016/j.compag.2024.109414

Walker J P, Willgoose G R & Kalma J D (2004). In situ measurement of soil moisture: A comparison of techniques. *Journal of Hydrology 293*(1–4): 85–99. https://doi.org/10.1016/j.jhydrol.2004.01.008

Wang R, Zhang J, Shan B, He M & Xu J (2022). XGBoost Machine Learning Algorithm for Prediction of Outcome in Aneurysmal Subarachnoid Hemorrhage. *Neuropsychiatric Disease and Treatment Volume* 18: 659–667. https://doi.org/10.2147/NDT.S349956

Wang Y, Shi L, Hu Y, Hu X, Song W & Wang L (2024). A comprehensive study of deep learning for soil moisture prediction. *Hydrology and Earth System Sciences* 28(4): 917–943. https://doi.org/10.5194/hess-28-917-2024

Xia Y, Jiang S, Meng L & Ju X (2024). XGBoost-B-GHM: An Ensemble Model with Feature Selection and GHM Loss Function Optimization for Credit Scoring. *Systems* 12(7): 254. https://doi.org/10.3390/systems12070254

Yetik A K, Arslan B & Şen B (2024). Trends and variability in precipitation across Turkey: A multimethod statistical analysis. *Theoretical and Applied Climatology* 155(1): 473–488. https://doi.org/10.1007/s00704-023-04645-4

Zeraatpisheh M, White A, Darby H, Neher D A, Faulkner J, Hancock L, Kretzler B, El-Naboulsi N, Turner H C, Von Wettberg E J, Wagner C H & Galford G L (2025). Spatial mapping and predictive modeling of soil organic carbon stocks in Vermont agricultural lands using machine learning and environmental variables. *Computers and Electronics in Agriculture* 237: 110727. https://doi.org/10.1016/j.compag.2025.110727

Zha X, Jia S, Han Y, Zhu W & Lv A (2025). Enhancing Soil Moisture Prediction in Drought-Prone Agricultural Regions Using Remote Sensing and Machine Learning Approaches. *Remote Sensing* 17(2): 181. https://doi.org/10.3390/rs17020181

Zhang Z, Zhang T & Li J (2023). Unbiased Gradient Boosting Decision Tree with Unbiased Feature Importance. Proceedings of IJCAI-23. https://www.ijcai.org/proceedings/2023/0515.pdf

Zhang T, Liang Z, Zhou J, Shao Q, Sarukkalige R, Lü H, Zhang J, Bi C, Wang J, Hu Y & Li B (2025). Multi-layer grid-scale soil moisture estimation using spatiotemporal deep learning methods with physical constraints. *Journal of Hydrology* 657: 133086. https://doi.org/10.1016/j.jhydrol.2025.133086

Zhu L, Dai W, Huang J & Luo Z (2025). A comparative analysis of deep learning models for accurate spatio-temporal soil moisture prediction. *Geocarto International* 40(1): 2441382. https://doi.org/10.1080/10106049.2024.2441382

Copyright © 2026 The Author(s). This is an open-access article published by Faculty of Agriculture, Ankara University under the terms of the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is properly cited.